Simulating 500 million years of evolution with a language model.

Year: 2025 | Citations: 134 | Authors: Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay

Abstract

More than three billion years of evolution have produced an image of biology encoded into the space of natural proteins. Here we show that language models trained at scale on evolutionary data can generate functional proteins that are far away from known proteins. We present ESM3, a frontier multimodal generative language model that reasons over the sequence, structure, and function of proteins. ESM3 can follow complex prompts combining its modalities and is highly responsive to alignment to improve its fidelity. We have prompted ESM3 to generate fluorescent proteins. Among the generations that we synthesized, we found a bright fluorescent protein at a far distance (58% sequence identity) from known fluorescent proteins, which we estimate is equivalent to simulating five hundred million years of evolution.